

ПРИМЕНЕНИЕ ГРИД-СЕРВИСОВ В ВЫЧИСЛИТЕЛЬНОЙ ХИМИИ: РЕАЛИЗАЦИЯ В ИПХФ РАН

**В.М. Волохов¹, Д.А. Варламов^{1,2}, А.В. Волохов¹, А.В. Пивушков¹, Г.А. Покатович¹,
Н.Ф. Сурков¹**

¹*Институт проблем химической физики РАН, Черноголовка*

²*Институт экспериментальной минералогии РАН, Черноголовка*

Введение

Вычислительная и квантовая химия являются одними из наиболее заинтересованных в крупномасштабных грид-вычислениях (в том числе на входящих в состав грид-полигонов суперкомпьютерах) отраслями науки. Исследования в области химии и смежных наук зачастую абсолютно неэффективны без использования сверхмощных параллельных и распределенных вычислительных ресурсов для решения задач разных классов. Например, некоторые задачи оптимизации высокомолекулярных структур требуют выполнения до 10^9 отдельных расчетов. Подобные расчеты требуют вычислительных ресурсов, которые не может предоставить ни один из вычислительных центров, что приводит к необходимости использования мощностей крупных распределенных грид-полигонов.

1. Ресурсный грид-центр ИПХФ РАН, структура и возможности

Работы с системами распределенных вычислений в ИПХФ РАН были начаты в 2004-2005 годах и ведутся до сих пор в рамках программ Президиума РАН, гранта РФФИ, Федеральных целевых программ, Программ Союзного Государства.

Основные направления работ: 1) адаптация прикладных пакетов в области вычислительной (квантовой) химии к работе в различных распределенных средах и обеспечение доступа широкого круга пользователей к работе с ними; 2) развитие ресурсного грид-центра (интегрирующего ресурсные сайты нескольких российских грид-полигонов) как в роли полигона для проведения вычислительных экспериментов, так и средства для решения реальных задач; 3) развитие новых методов вычислений и организации грид-сервисов в условиях распределенных сред.

Созданный в ИПХФ РАН ресурсный грид-центр объединяет в своем составе полнофункциональные ресурсные сайты следующих российских грид-полигонов:

- сайт Национальной Нанотехнологической Сети (ГридННС, <http://www.ngrid.ru>, виртуальная организация NanoChem) – среда Globus Toolkit 4, <http://www.globus.org>);
- сайт категории «А» СКИФ-Полигона (<http://skif-grid.botik.ru>) на базе промежуточного ПО Unicore (<http://www.unicore.eu>);
- узел консорциума EGI-RDIG (European Grid Infrastructure и Russian Data Intensive GRID, <http://www.egee-rdig.ru>) на основе среды gLite (<http://glite.web.cern.ch>), виртуальная организация (ВО) RGSTEST.

Данные ресурсные сайты позволяют решать *входящие* задачи с использованием адаптированных к распределенным средам прикладных квантово-химических пакетов и общего характера (не требующих предустановки на кластере).

В состав ресурсных сайтов входит также комплекс клиентских интерфейсов различных уровней для взаимодействия адаптированного квантово-химического прикладного ПО с грид-средами. Они позволяют запускать *исходящие* задачи вычислительной химии на распределенных ресурсах указанных полигонов, обеспечивая возможность формирования заданий, запуск на удаленных сайтах, мониторинга прохождения заданий, сбора результатов и статистики.

Важнейшей частью ресурсного центра ИПХФ РАН является грид-портал (<http://grid.icp.ac.ru>, Grid Enabled Chemical Physics и виртуальные сайты <https://nanochem.icp.ac.ru> и <https://webgrid.icp.ac.ru>), объединяющий высокоуровневые web-интерфейсы нескольких квантово-химических пакетов в систему грид-сервисов.

В настоящее время (сентябрь 2011 г.) ресурсный грид-центр ИПХФ реализован следующим образом. Установлен и настроен вычислительный кластер – на базе системы пакетной обработки заданий PBS/Torque Resource Manager подключены 6 расчетных узлов (с текущей пиковой производительностью до 172 Гфлопс). Управляющая и расчетная сеть – Gigabit Ethernet. Ресурсный сайт доступен для вычислений и мониторинга извне в постоянном режиме (24/7/365). На расчетных узлах установлен Scientific Linux 5.4 Boron, поскольку наличие данного дистрибутива требуется для расчетных узлов ресурсного сайта gLite, а для других распределенных сред выбор ОС (варианты Linux) не принципиален. На 2-х управляющих машинах установлен гипервизор KVM (Kernel-based Virtual Machine, <http://www.linux-kvm.org>). На управляющих машинах были установлены сервисные виртуальные машины:

- среда Globus Toolkit 4 (полигон ГридННЦ) – ОС CentOS 5.4, сервисы MDS, GRAM, GridFTP, RFT, User Interface;
- среда Unicore 6.2 (СКИФ-Полигон) – ОС Ubuntu 9.10, сервисы шлюз (Gateway); серверный контейнер (Unicore/X), интерфейс к целевой системе (TSI), сервис авторизации и пользовательская база данных XUUDBS, клиентский интерфейс (UI);
- среда gLite 3.1 (полигон EGEE-RDIG) – ОС ScientificLinux 4.5, Computing Element и соответствующие сервисы типа lcg-ce, сервисов авторизации и мониторинга.

На все управляющие машины установлены серверные компоненты пакета управления заданиями PBS/Torque (<http://www.clusterresources.com>), на расчетные узлы установлена клиентская часть данного пакета. Поскольку все распределенные middleware требуют наличия своих собственных очередей PBS, было принято решение о настройке трех одновременно работающих экземпляров сервиса pbs_tom на расчетных узлах с соответствующим набором очередей заданий. В таком варианте каждая управляющая машина связывается с расчетными узлами по уникальному порту и имеет дело только со своими заданиями. Недостатком данного подхода является невозможность (пока) полного учета ресурсов, используемых расчетными узлами, но для вычислительных экспериментов и отладки прикладного ПО это вполне приемлемо.

2. Основные варианты работы с грид-сервисами для решения задач вычислительной химии

Работа в распределенных средах с прикладными пакетами вычислительной химии и авторскими программами

Авторами проведена адаптация ряда стандартных прикладных пакетов, используемых в вычислительной химии, и различных авторских программ, разработанных в ИПХФ и ИЦЧ РАН для проведения распределенных вычислений на максимуме доступных ресурсов российских и международных грид-инфраструктур.

Для адаптации в вышеупомянутых грид-средах были выбраны следующие прикладные квантово-химические программные пакеты:

- **GAMESS-US** (<http://www.msg.ameslab.gov/GAMESS>) – расчет энергий, структур молекул, частот их колебаний, а также разнообразные свойства молекул в газовой фазе и в растворе, как в основном, так и в возбужденных состояниях.
- **VASP** (<http://cms.mpi.univie.ac.at/vasp>) и **PWscf** (<http://www.pwscf.org>) – необходимы для моделирования процессов на поверхности и в объеме твердых тел (прежде всего катализа и ионной проводимости). Моделирование на квантово-механическом уровне осуществляется для малых кластеров с числом атомов 10-100, определяющих существование возможных в материале фаз.
- **Gaussian-03** (<http://www.gaussian.com>) – позволяет рассчитывать энергию, структуру молекул, частоты их колебаний, а также разнообразные свойства молекул в газовой фазе и в растворе, в основном и в возбужденных состояниях, включая методы расчета сверхбольших молекулярных систем (использование пакета в распределенных средах затруднено лицензионными ограничениями).
- **NAMD** (<http://www.ks.uiuc.edu/Research/namd>) – обладает высокой эффективностью распараллеливания, используется для моделирования больших молекулярных систем (до миллионов атомов, вплоть до расчета ДНК структур).
- **NWChem 6.0** (<http://www.nwchem-sw.org>) – реализация смешанных стратегий квантово-механического и молекулярно-динамического моделирования, в которых определенные локальные области материала моделируются на квантово-механическом, а другие – на молекулярно-динамическом уровнях, используется в области кинетики и динамики химических превращений, химии на границах фаз и в конденсированных (твердых и жидких) фазах.
- **Dalton-2** (<http://www.kjemi.uio.no/software/dalton/dalton.htm>) – расчет синглет-синглетных возбуждений, электронных структур, вращательных и колебательных спектров молекул, учет релятивистских эффектов и эффектов сольватации.
- **CPMD** (<http://www.cpmd.org>) – расчеты в области молекулярной динамики.
- **Авторские программы** (разработки ИПХФ), включающие многопараметрические задачи из области квантовой химии и молекулярной динамики.

Для большинства выбранных прикладных пакетов созданы и протестированы на реальных задачах низкоуровневые интерфейсы для запуска их в распределенных вычислительных средах. Данные интерфейсы включают набор скриптов по формированию исходящих заданий, запуску через грид-инфраструктуру на удаленных ресурсных узлах, мониторингу выполнения задач, возвращению полученных результатов и «сборку» окончательных результатов на интерфейсе пользователя. На грид-центре ИПХФ, использованном в качестве удаленного распределенного ресурса, проведены запуски указанного прикладного ПО через инфраструктуры ГридННС СКИФ-Полигона, EGI-RDIG. Запуски всего адаптированного ПО проводились в разных режимах и конфигурациях (с разным количеством процессоров и использованием разных вариантов параллельных расчетов). В настоящее время ресурсные сайты ИПХФ позволяют решать в качестве входящих задачи вычислительной химии с использованием всех указанных квантово-химических пакетов.

Реализация грид-сервисов в виде высокоуровневых web-интерфейсов

Составной частью ресурсного центра ИПХФ РАН является грид-портал, объединяющий грид- и web-сервисы. В его рамках сформированы высокоуровневые web-интерфейсы, позволяющие более эффективно использовать все преимущества грид-расчетов. Эта среда позволяет пользователям получить доступ к грид-ресурсам и сервисам, вызывать и настраивать их с помощью web-браузера. Архитектура грид-портала основана на идее, что порталная система является контейнером для низкоуровневых пользовательских интерфейсов, обеспечивающих работу с грид-службами. Преимущество

данной архитектуры в том, что она достаточно легко позволяет встраивать в портал интерфейсы новых грид-служб и изменять существующие. Портальные сервисы контролируют и визуализируют пользовательский интерфейс. В ИПХФ РАН сформирован грид-портал (<http://grid.icp.ac.ru>, Grid Enabled Chemical Physics и виртуальные защищенные сайты), включающий www-интерфейсы к следующим прикладным пакетам:

1. GAMESS-US, Gaussian, NAMD, PWscf, NWChem;
2. комплекс по вычислению многопараметрических функций для задач химической физики, обладающих свойством параллелизма по данным (Data Parallel).

Данные web-интерфейсы позволяют определять входные параметры и условия (включая загрузку данных, создание и редакцию конфигурационных файлов, работу с сертификатами пользователя), формировать сложные первичные файлы запуска, производить запуск данного ПО в распределенных средах, осуществлять мониторинг выполнения заданий и сбор результатов. Интегрирована также технология работы через web-интерфейс с «пучками» независимых заданий на «нарезаемых» областях данных, а также с использованием метода формирования «виртуальных контейнеров» (для пакета GAMESS-US). Заметим, что основная часть программного кода web-интерфейсов не связана напрямую с выбранной распределенной средой, поэтому они могут быть подключены ко всем трем вышеупомянутым грид-полигонам (EGI-RDIG, СКИФ-Полигон, ГридННЦ) и, соответственно, поддерживают три распределенные среды (gLite, Unicore, Globus GT4).

Созданные web-интерфейсы значительно снижают трудоемкость работы пользователя в части формирования задач и работы с первичными данными и значительно облегчают последующую работу с пакетами в распределенных средах.

Решение грид-задач с применением методов «виртуальных контейнеров»

Большинство прикладных пакетов отличаются сложностью конфигураций и повышенными требованиями к среде выполнения, особенно для проведения параллельных расчетов. Обычно эта проблема решается путем создания виртуальных организаций, т.е. объединением через распределенные среды во многом однотипных (по установленному ПО и настройкам) вычислительных ресурсов. Для них прикладные пакеты (вместе со средствами конфигурирования и настройки) распространяются из единого репозитория. В большинстве же случаев неподготовленный ресурсный сайт не имеет нужного заранее установленного прикладного ПО или хотя бы не сконфигурирован должным образом, поэтому запуск непредустановленных сложных прикладных пакетов обычно для таких ресурсов оканчивается неудачей. В общем случае необходима ручная или полуавтоматическая перенастройка ресурсных узлов распределенных сред, включающая установку пакетов, конфигурирование центрального узла и расчетных узлов (настройка переменных окружения, общих NFS ресурсов, PBS очередей), установка дополнительных системных библиотек и исполняемых файлов (включая параллельные среды типа Mpi-2). При условии этого возможны запуски пакетов на распределенных ресурсах (как это и сделано на ресурсных сайтах ИПХФ для решения входящих задач).

Для частичного решения данной проблемы авторами разработан метод создания виртуальных перемещаемых программных «контейнеров». «Контейнер», включающий собственно прикладной пакет, набор необходимых системных файлов и библиотек, скрипты по развертыванию и настройке среды исполнения, файлы данных и конфигурационные файлы, доставляется на удаленный ресурсный узел грид-среды стандартными средствами распределенного middleware. Применение таких «контейнеров» позволяет передавать заранее настроенную среду как единое задание, не требующее дополнительного конфигурирования и сложной процедуры установки и настройки, производимых, как правило, вручную администратором кластеров. «Контейнер» по прибытии на ресурсный узел проводит развертывание пакета и необходимых системных библиотек.

лиотек, настройку среды исполнения (включая параллельную среду), запуск задания, по его окончании проводится отправка результатов на пользовательский интерфейс и «очистка» среды исполнения, т.е. приведение ресурса в первоначальное состояние. Так могут быть решены задачи установки, настройки, несовместимости с операционной системой и другими программами, разрешаются конфликты одинаковых приложений.

Работа с «пучками» формально независимых заданий

Для решения широкого класса многопараметрических задач вычислительной химии с использованием грид-технологий был создан метод запуска «пучков» независимых заданий для использования всех доступных ресурсов распределенной среды. В области химической физики существует класс задач, требующих перебора большого количества независимых параметров. При этом полная задача разбивается на огромное количество независимых подзадач (каждая определяется группой значений совокупности параметров). Подобные задачи имеют высокую вычислительную сложность, однако вычисления в точках сетки в них происходят независимо друг от друга, поэтому достаточно разбить область вычислений на множество подобластей и для каждой из них запускать задачу на различных процессорах.

Разработана методика запуска задач и получения результатов методом запуска «пучков» заданий на всех доступных ресурсах выбранной распределенной среды. На языке Perl написан комплекс программ для запуска «пучков» заданий и получения результатов счета с использованием пользовательских интерфейсов (UI) сред gLite, Unicore, Globus GT4. Для решения многопараметрических задач квантовой химии были разработаны методы формирования «пучков» независимых заданий с варьирующими параметрами – до 10^4 , в перспективе до 10^7 «атомарных» заданий на задачу. Для выбранных областей данных авторскими скриптами проводится «нарезка» областей данных, формирование пулов независимых заданий, создание очередей запуска и отправки заданий на брокер ресурсов. После запуска заданий периодически запускаемые средствами ОС (например, по *cron*) скрипты ведут мониторинг выполнения, контроль таймаутов, перезапуск неудачных заданий и сбор результатов выполненных заданий (с использованием SQL базы данных и таблиц в ней, контролирующих состояние заданий – «ожидание», «запуск», «выполнение» и т.д.). По окончании расчетов проводится сборка «атомарных» результатов в единый выходной файл.

Главным недостатком механизма работы с «пучками» являются ограниченные возможности пользователя по мониторингу собственных заданий в средах Unicore и Globus GT4. В целом же, описанный новый, полностью асинхронный механизм запуска «пучка» заданий требует от пользователя только выбора ресурса по желанию, что возможно через механизмы web-портала.

Заключение

В ИПХФ РАН создан грид-центр, позволяющий проводить масштабные расчеты в области вычислительной химии в распределенных средах на крупномасштабных полигонах (в перспективе до 10^4 CPU). Продемонстрирована применимость созданных ресурсов для решения крупномасштабных химических задач на высокопроизводительных вычислительных полигонах. Это позволяет ставить и решать вычислительные задачи фундаментального и прикладного характера в области химических наук, ранее не доступные из-за ограниченности возможностей вычислительных ресурсов. Основные научные области применения – химическая физика, квантовая химия, исследование наноструктур, молекулярная динамика, биохимия, фармацевтика, разработка топливных элементов и близкие отрасли наук.

Работа выполнена при поддержке гранта РФФИ № 11-07-00686-а.